УДК 004.93

Л.Г. Ахметшина, А.А. Егоров, И.М. Удовик
Днепропетровский национальный университет имени Олеся Гончара, Украина
Украина, 49050, г. Днепропетровск, ул. Научная, 13

Улучшение сходимости нейро-фаззи кластеризации многомерных данных при использовании неевклидовых метрик

L.G. Achmetshina, A.A. Yegorov, I.M. Udovick
Dniepropetrovsk National University named by Oles Hochar, Ukraine
Ukraine, 49050, c. Dniepropetrovsk, Nauchnaja av., 13

The Sensitivity of The Neuro-fuzzy Clustering Improvement Based On Non-Euclidian Metrics

Л.Г. Ахметшина, А.О. Єгоров, І.М. Удовик
Дніпропетровський національний університет імені Олеся Гончара, Україна
Україна, 49050, г. Дніпропетровськ, вул. Наукова, 13

Підвищення чутливості нейро-фаззі кластеризації багатовимірних даних на основі неевклідових метрик

В статье предложен модифицированный алгоритм гибридной нечеткой кластеризации mdsFCM, который благодаря применению матрицы расстояний Махаланобиса в процессе подготовки центроидов к обработке сетью Кохонена и выполнения сжатия ее размера, позволяет повысить сходимость и, в ряде случаев, чувствительность при обработке многомерных данных. Представлены экспериментальные результаты применения предложенного модифицированного алгоритма mdsFCM для кластеризации низкоконтрастных цветных медицинских изображений.
Ключевые слова: многомерные изображения, нейро-фаззи кластеризация, меры расстояний, сегментация, неевклидовые метрики.

This article deals with the description of the hybrid fuzzy clustering algorithm mdsFCM which is used non-Euclidian distances based on calculation the covariance matrix. This algorithm has the greater level of sensitivity while processing multidimensional data. The experimental results of the application of proposed algorithm for low-contrast medical color images clustering are shown.
Keywords: multidimensional images, neuro-fuzzy clustering, distance measures, segmentation, non-Euclidean metrics.

В статті запропоновано алгоритм гібридної нечіткої кластеризації mdsFCM, який завдяки застосуванню неевклідових метрик заснованих на використанні матриці коваріації, має більш високий рівень чутливості при обробці багатовимірних даних. Представлені експериментальні результати застосування запропонованого алгоритму для кластеризації низькоконтрастних кольорових медичних зображень.
Ключові слова: багатовимірні зображення, нейро-фаззі кластеризація, міри відстані, сегментація, неевклідові метрики.

Введение

Кластеризация является одной из часто решаемых задач, возникающих при обработке и анализе данных. И хотя на сегодняшний день разработано более 200 различных методов выполнения кластеризации их эффективность существенно меняется в зависимости от решаемой задачи и специфики входных данных.

Достаточно часто для решения задачи кластеризации применяются нечеткие или нейросетевые алгоритмы. Одним из самых распространенных нечетких алгоритмов является FCM (Fuzzy c-means) [1], что обусловлено его простотой и достаточной для многих задач чувствительностью. Однако этот метод не учитывает, специфику объектов, имеющих множество информативных признаков, в отличие от алгоритмов Густафсона-Кесселя и FMLE (Fuzzy Maximum Length Estimates) [2], которые используют неевклидовые метрики, основанные на вычислении матрицы ковариации, что позволяет повысить достоверность анализа многомерных данных.

Среди нейросетевых алгоритмов интерес представляют сети, обучающиеся без учителя, примером которой является самоорганизующаяся карта Кохонена (SOM), отличающаяся простотой архитектуры, относительно высоким быстродействием и не требующая длительной процедуры настройки весов [3].

Популярным в настоящее время также является нейро-фаззи технология выполнения кластеризации, предполагающая объединение нечеткого алгоритма и нейронной сети в пределах одного метода для повышения чувствительности или быстродействия. В работе [4] был предложен метод гибридной нечеткой кластеризации многомерных данных mdsFCM, являющийся примером реализации такой технологии. Этот метод представляет собой объединение нечеткого алгоритма, использующего неевклидовые метрики, и SOM, применяемой на каждой итерации для уточнения значений центроидов с целью повышения чувствительности.

Постановка задачи

Для повышения чувствительности в методе mdsFCM перед применением SOM осуществляется формирование дополнительных значений центроидов на основе применения метода пропорционального распределения [5]. Соответственно, после выполнения кластеризации сетью Кохонена, в процессе которой осуществляется динамическое уменьшение ее размера, вычисляются значения центроидов текущей итерации на основе выбора наиболее значимых нейронов.

Целью данной статьи является улучшение сходимости и достоверности метода гибридной нечеткой кластеризации многомерных данных mdsFCM за счет применения матрицы расстояний Махаланобиса в процессе увеличения числа центроидов, сжатия размерности карты Кохонена и выбора наиболее значимых нейронов при формировании центров кластеров текущей итерации.

Решение задачи

Предложенный модифицированный метод гибридной нечеткой кластеризации многомерных данных mdsFCM состоит из двенадцати шагов.

1. Инициализация начальных значений числа нечетких кластеров c и центроидов v_{fcm}^{0}; экспоненциального веса нечеткой кластеризации m; коэффициента увеличения числа нечетких кластеров N_e.

2. Формирование начальных значений векторов весов нейронов SOM на основании матрицы центроидов предыдущей итерации v_{fcm}^{t-1}, количество которых выбирается равной $[N_e \cdot c, 1]$ (одномерная структура – столбец). Дополнительные значения v_{fcm}^{t-1} вычисляется методом пропорционального распределения – формирования для каждой пары упорядоченных центров v_{fcm}^{t-1} новых центроидов, количество которых пропорционально расстоянию Махаланобиса между ними.

3. Кластеризация исходных данных с помощью SOM, происходящая в два этапа: грубая и тонкая настройка весов нейронов. При этом на каждой итерации выполняется динамическое уменьшение размерности сети, основанное на использовании матрицы расстояний Махаланобиса между нейронами, которая рассчитывается с применением их весов.

4. Получение новых значений центров нечетких кластеров v_{som}^t путем выбора c значимых центров из матрицы весов нейронов, полученной в результате обучения SOM. В процессе этого выбора применяется матрица расстояний Махаланобиса между нейронами.

5. Вычисление текущих значений функции принадлежности u^t:

$$u_{k,i}^t = \sum_{L=1}^{c} \left[\frac{D_{i,k}}{D_{i,L}} \right]^{\frac{-2}{m-1}} \begin{pmatrix} \forall k \in \{1,...,c\}, \\ \forall i \in \{1,...,n\} \end{pmatrix}, \tag{1}$$

где n – число экземпляров данных, а D – матрица расстояний между экземплярами исходных данных X и центрами нечетких кластеров, которая вычисляется по следующей формуле:

$$D_{i,k} = \sqrt{\left(X_i - \left(v_{som}^t\right)_k\right)^T \cdot A \cdot \left(X_i - \left(v_{som}^t\right)_k\right)}, \left(\forall i \in \{1,...,n\}, \forall k \in \{1,...,c\}\right), \tag{2}$$

причем $A = A_k^t, \left(\forall k \in \{1,...,c\}\right)$. Следует отметить, что способ формирования матриц A_k^t влияет на чувствительность кластеризации.

6. Автоматическое определение количества нечетких кластеров благодаря динамическому сжатию функции принадлежностей, на основе расстояний между центрами нечетких кластеров.

7. Формирование матрицы центров нечетких кластеров v_{fcm}^t, которые будут использованы в начале следующей итерации:

$$(v_{fcm}^t)_{k,j} = \left(\sum_{i=1}^{n} \left(u_{k,i}\right)^m \cdot X_{i,j} \right) / \sum_{i=1}^{n} \left(u_{k,i}\right)^m . \tag{3}$$

8. Вычисление значения Δ_v^t – среднего по матрице расстояний между центрами нечетких кластеров v_{fcm}^t и v_{fcm}^{t-1}, а также критериев V_{xb}^t и V_{fz}^t, которые являются показателями Ксие-Биени и нечеткости текущей итерации, соответственно, следующим образом [5]:

$$V_{xb}^t = \left(\sum_{k=1}^{c} \sum_{i=1}^{n} \left(u_{k,i}^t\right)^m \cdot \sum_{j=1}^{q} \left(X_{i,j} - \left(v_{fcm}^t\right)_{k,j} \right)^2 \right) / \left(n \cdot \left(d_{min}\right)^2 \right), \tag{4}$$

$$V_{fz}^t = \left(\sum_{k=1}^{c} \sum_{i=1}^{n} \left(u_{k,i}^t\right)^2 \right) / n, \tag{5}$$

где d_{min} – минимальное Евклидово расстояние между центрами нечетких кластеров.

9. Если выполняется условие $C_{fz}^t \ge C_{fz}^{max}$, причем $C_{fz}^t = V_{fz}^t / V_{xb}^t$, а C_{fz}^{max} – максимальный из коэффициентов C_{fz}^t, то запоминаются следующие значения: $\Delta_v^{max} = \Delta_v^t$, $C_{fz}^{max} = C_{fz}^t$, $u^{max} = u^t$ и $v_{fcm}^{\max} = v_{fcm}^t$..

10.Если разность $\Delta_v^t - \Delta_v^{t-1}, (\forall t > 1)$ меняет свой знак, 2, 4, 8 и т.д. раз, то пороговое значение ε увеличивается в 10 раз.

11.Если не выполняются условия:

$$\Delta_v^t < \varepsilon \text{ или } \left(\left| V_{xb}^t - V_{xb}^{t-1} \right| < \varepsilon \text{ и } \left| V_{fz}^t - V_{fz}^{t-1} \right| < \varepsilon \right), \tag{6}$$

$$\left(\sum_{k=1}^{c} \sqrt{\sum_{j=1}^{q} \left(\left(v_{fcm}^t \right)_{k,j} - \left(v_{fcm}^{t^1} \right)_{k,j} \right)^2} \right) / c < \varepsilon, \left(\forall t^1 \in \{t-1,...,1\} \right), \tag{7}$$

где V_{xb}^{t-1} и V_{fz}^{t-1} – показатели Ксие-Биени и нечеткости предыдущей итераций, соответственно, то осуществляется переход к пункту 2.

12.Если выполняется условие:

$$C_{fz}^t < C_{fz}^{max} \text{ и } \left(\Delta_v^t > \Delta_v^{max} \text{ или } \left(\Delta_v^t < \Delta_v^{max} \text{ и } p_{\Delta_v} > p_c \right) \right), \tag{8}$$

причем

$$p_c = \frac{\left| C_{fz}^t - C_{fz}^{max} \right|}{max(C_{fz}^t, C_{fz}^{max})} \cdot \frac{1}{C_{fz}^{max} - C_{fz}^{min}}, \tag{9}$$

$$p_{\Delta_v} = \frac{\left| \Delta_v^t - \Delta_v^{max} \right|}{max(\Delta_v^t, \Delta_v^{max})} \cdot \frac{1}{\left(\Delta_v^{max} \right)' - \Delta_v^{min}}, \tag{10}$$

где C_{fz}^{min} и Δ_v^{min} – минимальные значения параметров C_{fz}^t и Δ_v^t, соответственно, а $\left(\Delta_v^{max} \right)'$ – максимальное значение критерия Δ_v^t, то происходит возврат к сохраненным значениям матриц нечеткой функции принадлежностей u^{max} и центров нечетких кластеров v_{fcm}^{max}, которые и являются результатом обучения.

В данной работе A_k^t перед применением формулы (2) формировались с использованием нечеткой матрицы ковариации по формуле:

$$A_k^t = \left(\rho_k * det\left(F_k^t \right) \right)^{1/q} * \left(F_k^t \right)^{-1}, \tag{11}$$

где ρ_k – константа, отражающая знания о данных подлежащих группированию (если таких знаний до начала кластеризации нет, то $\rho_k = 1, (\forall k \in \{1,...,c\})$), а F_k^t – так называемая нечеткая матрица ковариации *k*-й группы – формируется следующим образом:

$$F_k^t = \frac{\sum_{i=1}^{n} \left(\left(u_{som}^t \right)_{i,k} \right)^m \left(X_i - \left(v_{som}^t \right)_k \right) \left(X_i - \left(v_{som}^t \right)_k \right)^T}{\sum_{i=1}^{n} \left(\left(u_{som}^t \right)_{i,k} \right)^m}, \left(\forall k \in \{1,...,c\} \right), \tag{12}$$

причем u_{som}^t вычисляется по формуле (1), при использовании которой формирование матриц F_k^t, необходимых для получения A_k^t, осуществляется по формуле:

$$F_k^t = \sum_{i=1}^{n} \left(X_i - \left(v_{som}^t \right)_k \right) \left(X_i - \left(v_{som}^t \right)_k \right)^T, \left(\forall k \in \{1,...,c\} \right). \tag{13}$$

При этом u^0 вычисляется по формуле (2) с использованием матриц v^0 и $A = I$;

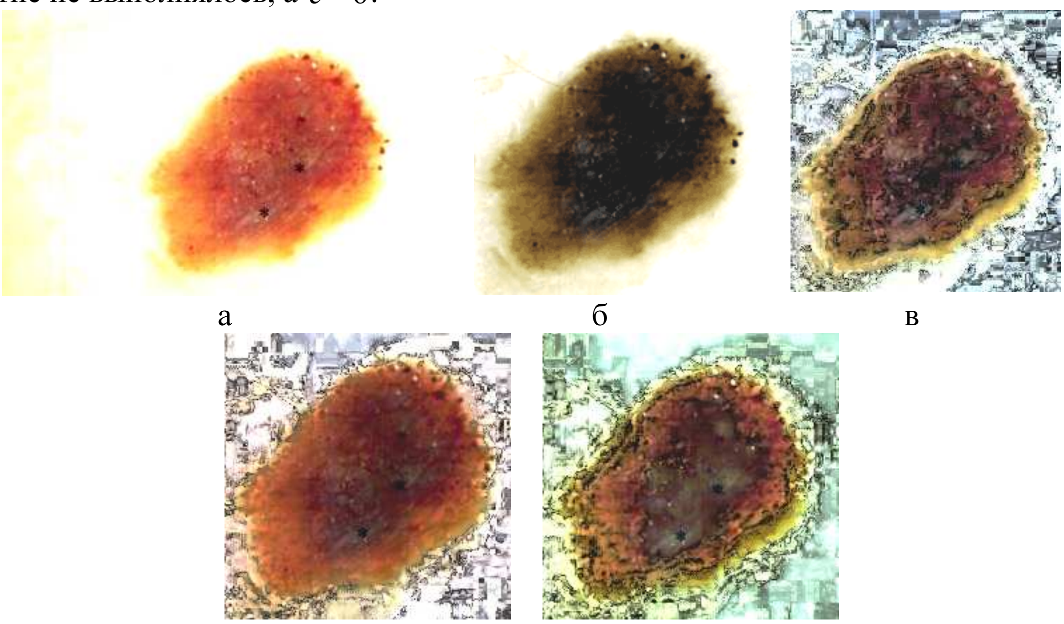
Экспериментальные результаты были получены при обработке различных цветных низкоконтрастных изображений, в том числе медицинских, примером которых служат RGB снимки, приведенные на рис. 1 а и 2 а, представляющие собой результаты дерматоскопии с целью диагностирования меланомы по визуальным признакам, среди которых основными являются наличие бело-синих структур и пятен неправильной формы.

При кластеризации использовались следующие значения управляющих параметров: $N_e = 5$ (рекомендуемые значения – 5 или 6, а допустимые значения – целые числа от 2 до 8); после применения карты Кохонена выбирались центры кластеров на основе максимума показателя обоснованности кластера [6]. Суть метода заключается в последовательном выборе c нейронов, для которых значение показателя Val_k обоснованности кластера [7] максимально. Показатель Val_k для каждого нейрона рассчитывается следующим образом:

$$Val_k = \sum_{i=1}^{n} \left(u_{k,i} \cdot \left(d_{k,i} \right)^2 \right), \left(\forall k \in \{1,..., N \cdot c\}, \forall i \in \{1,..., n\} \right), \tag{14}$$

где $d_{k,i}$ – Евклидово расстояние от центра k-го нейрона до каждого экземпляра исходных данных. Визуализация результатов нечеткой кластеризации производилась методом сравнения с исходными данными на основе максимального соответствия [8]. Размерность карты Кохонена выбиралась равно 16×14 нейронов.

При обработке изображения на рис. 1 а методом mdsFCM применялось динамическое сжатие нечеткой функции принадлежностей (на основе матрицы расстояний Махаланобиса), причем $c = 20$. При кластеризации снимка на рис. 2 а динамическое сжатие не выполнялось, а $c = 6$.



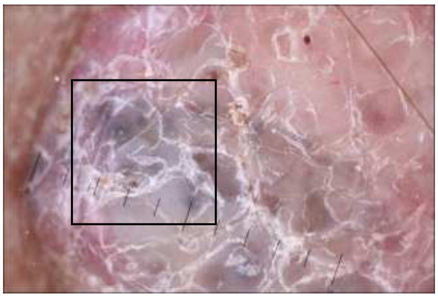
а б в

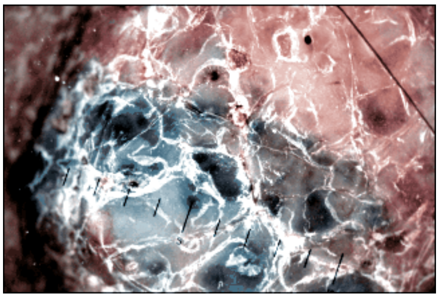
г д

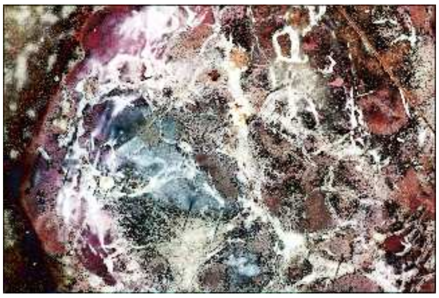
Рисунок 1 – Кластеризация цветного медицинского изображения:
а – исходный снимок (330x214 пикселей); фрагмент результатов обработки методами: б – SOM; в – Густафсона-Кесселя; г – mdsFCM;
д – модифицированный mdsFCM

На изображении, представленном на рис. 1 а, звездочками отмечены области, содержащие очаги меланомы. Применение SOM (рис. 1 б) и методов Густафсона-Кесселя (рис. 1 в) и mdsFCM (рис. 1 г) не приводит к их выделению, в то время как использование модифицированного алгоритма mdsFCM (рис. 1 д) позволяет справиться с этой задачей. Кроме того, использование модифицированного алгоритма mdsFCM позволило сократить число итераций на 11%. При этом в результате сжатия было получено 20 (рис. 1 г) и 12 (рис. 1 д) кластеров, соответственно.

На изображении, представленном на рис. 2 а, в области интереса, обведенной прямоугольником, содержатся плохо различимые на исходном снимке бело-синие структуры, что затрудняет визуальное диагностирование как самой меланомы, так и области ее распространения. Использование SOM (рис. 2 б) привело к нарушению цветового баланса из-за чрезмерного заполнения синим цветом области интереса, что затрудняет адекватное визуальное диагностирование. Кластеризация методом Густафсона-Кесселя (рис. 2 г) не приводит к выделению области распространения меланомы. Применение как исходного, так и модифицированного алгоритмов mdsFCM (рис. 2 г, д) позволяет выполнить визуальное диагностирование, однако, в последнем случае за счет изменения цветопередачи удается точнее выявить область распространения меланомы. При этом применение модифицированного алгоритма mdsFCM позволило на 11% сократить число итераций.


а


б


в


г


д

Рисунок 2 – Кластеризация цветного медицинского изображения:
а – исходный снимок (323x215 пикселей); результаты обработки методами:
б – SOM; в – Густафсона-Кесселя; г – mdsFCM; д – модифицированный mdsFCM

Следует отметить, что при проведении экспериментов модифицированный алгоритм гибридной нечеткой кластеризации mdsFCM позволил в среднем на 17.4% сократить число итераций обучения по сравнению с исходным методом.

Выводы

Предложенный в данной работе модифицированный метод гибридной нечеткой кластеризации многомерных данных mdsFCM позволяет улучшить сходимость и, в ряде случаев, достоверность кластеризации цветных изображений по сравнению с исходным алгоритмом. При этом улучшение сходимости, в среднем, более заметно при выполнении динамического сжатия нечеткой функции принадлежностей, а также зависит от метода выбора значимых нейронов после обучения карты Кохонена.

Литература

1. Леоненков А. Нечеткое моделирование в среде MATLAB и fuzzyTECH / А. Леоненков. – СПб. : БХВ-Петербург, 2003. – 719 с.
2. Рутковский Л. Методы и технологии искусственного интеллекта / Л. Рутковский. – М. : Горячая Линия-Телеком, 2010. – 600 с.
3. Кохонен Т. Самоорганизующиеся карты / Т. Кохонен ; [пер. 3-го англ. изд. В.Н. Агеева под ред. Ю.В. Тюменцева]. – М. : Бином. Лаборатория знаний, 2008. – 665 с.
4. Ахметшина Л.Г. Влияние вида меры расстояния на чувствительность нейро-фаззи кластеризации многомерных данных / Л.Г. Ахметшина, А.А. Егоров // Искусственный интеллект. – 2012. – № 4. – С. 535-545.
5. Ахметшина Л.Г. Повышение чувствительности гибридной нечеткой кластеризации на основе формирования центроидов пропорционально расстояниям в q-мерном пространстве / Л.Г. Ахметшина, А.А. Егоров // Геометричне та комп'ютерне моделювання. – 2009. – Вип. 24. – С. 193-198.
6. Ахметшина Л.Г. Влияние способов получения новых центроидов и выбора существенных кластеров на чувствительность алгоритма гибридной нечеткой кластеризации / Л.Г. Ахметшина, А.А. Егоров // Интеллектуальные системы принятия решений и проблемы вычислительного интеллекта: международная научная конф., 17 – 21 мая 2010 г., Евпатория. – Т. 1. – С. 231-234.
7. Пегат А. Нечеткое моделирование и управление / А. Пегат; [пер. с англ. А.Г. Подвесовского, Ю.В. Тюменцева]; под. ред. Ю.В. Тюменцева – М. : БИНОМ, 2009. – 768 с.
8. Егоров А.А. Визуализация результатов нечеткой кластеризации цветных изображений на основе метода сравнения с исходными данными // Егоров А.А. / Вестник Херсонского национального технического университета. – 2009. – № 2(35). – С. 195-199.

Literatura

1. Leonenkov A. Fuzzy modeling in the MATLAB and fuzzyTECH environment – S.P.: BHV–Peterburg. – 2003. – 719.
2. Rutkovsky L. The methods and technology of artificial intelligence – M.: Gorjachaja-Linija-Telekom, 2010. – 600.
3. Kohonen T. Self Organized Maps [transl. 3 engl. publ. V.N. Ageeva edited by J.V. Tumentseva]. – M.: Binom. Knowledge Laboratory – 2008. – 665.
4. Akhmetshina L.G., Yegorov A.A. The Influence of the distance measure type on the sensitivity neuron-fuzzy clustering of the multidimensional data. Artificial Intelligence. – 2012. – № 4. – S. 535 – 545.
5. Akhmetshina L.G., Yegorov A.A. Enhancement of the hybrid fuzzy clustering sensitivity based on forming of the centers proportional to the q-dimensional space distances. Geometric and computer modeling. – 2009. – Num. 24. – S. 193 – 198.
6. Akhmetshina L.G., Yegorov A.A. The influence of the new centers forming and essential clster choosing on the hybrid fuzzy clustering sensitivity // Intellectual systems for decision making and problems of computational intelligence: international science conf. May 17-21, 2010, Yevpatorija. – Vol. 1. – S. 231 – 234.

7. Pegat A. Fuzzy modeling and control. [transl. from engl. A.G. Podvesovskogo, Y.V. Tyumenceva]; Edited by. Y.V. Tyumenceva – M.: BINOM – 2009. – 768.
8. Yegorov A.A. The Fuzzy clustering results visualizing based on the initial data compare method. Bulletin of The Herson National Technical University. – 2009. – Num. 2(35). – S. 195 – 199.

RESUME

L.G. Achmetshina, A.A. Yegorov, I.M. Udovick

The Sensitivity Of The Neuro-fuzzy Clustering Improvement Based On Non-Euclidian Metrics

This article deals with the description and experimental verification of the modified algorithm of the hybrid fuzzy clustering of the multidimensional data – mdsFCM. The base mdsFCM algorithm [4] corresponds the union of the fuzzy clustering algorithm which used the non-Euclidian distances based on calculation of the covariance matrix and SOM (Self Organized Map). The essence of the algorithm modification consists in the using of the non-Euclidian distances during clustering by SOM. This leads to the algorithm convergence and sensitivity improvement for multidimensional data processing.

The experimental test of the proposed clustering method showed the convergence and, in some cases, sensitivity improvement in comparison with mdsFCM, sFCM, FCM (when non-Euclidian distances based on covariance matrix calculation are used) and Gustafson-Kessel algorithms. In the capacity of the initial data is used the various low-contrast color medical images which are the particular case of the multidimensional data. Experimental tests show that algorithm convergence improvement is depended on choosing most important neurons method after clustering by SOM.

Статья поступила в редакцию 26.04.2013.